Article

Identifying Kinematic Structures in Simulated Galaxies Using Unsupervised Machine Learning

Du, Min, Ho, Luis C., Zhao, Dongyao, Shi, Jingjing, Debattista, Victor P, Hernquist, Lars and Nelson, Dylan

Du, Min, Ho, Luis C., Zhao, Dongyao, Shi, Jingjing, Debattista, Victor P orcid iconORCID: 0000-0001-7902-0116, Hernquist, Lars and Nelson, Dylan (2019) Identifying Kinematic Structures in Simulated Galaxies Using Unsupervised Machine Learning. The Astrophysical Journal, 884 (129). ISSN 0004-637X

The Astrophysical Journal, 884:129 (11pp), 2019 October 20 https://doi.org/10.3847/1538-4357/ab43cc

Identifying Kinematic Structures in Simulated Galaxies Using Unsupervised Machine Learning

Min Du[1], Luis C. Ho[1,2], Dongyao Zhao[1], Jingjing Shi[1], Victor P. Debattista[3], Lars Hernquist[4], and Dylan Nelson[5]
[1] Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, People's Republic of China; dumin@pku.edu.cn
[2] Department of Astronomy, School of Physics, Peking University, Beijing 100871, People's Republic of China
[3] Jeremiah Horrocks Institute, University of Central Lancashire, Preston PR1 2HE, UK
[4] Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[5] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85741 Garching, Germany
Received 2019 May 3; revised 2019 August 30; accepted 2019 September 9; published 2019 October 18

Abstract

Galaxies host a wide array of internal stellar components, which need to be decomposed accurately in order to understand their formation and evolution. While significant progress has been made with recent integral-field spectroscopic surveys of nearby galaxies, much can be learned from analyzing the large sets of realistic galaxies now available through state-of-the-art hydrodynamical cosmological simulations. We present an unsupervised machine-learning algorithm, named auto-GMM, based on Gaussian mixture models, to isolate intrinsic structures in simulated galaxies based on their kinematic phase space. For each galaxy, the number of Gaussian components allowed by the data is determined through a modified Bayesian information criterion. We test our method by applying it to prototype galaxies selected from the cosmological simulation `IllustrisTNG`. Our method can effectively decompose most galactic structures. The intrinsic structures of simulated galaxies can be inferred statistically by non-human supervised identification of galaxy structures. We successfully identify four kinds of intrinsic structures: cold disks, warm disks, bulges, and halos. Our method fails for barred galaxies because of the complex kinematics of particles moving on bar orbits.

Key words: galaxies: fundamental parameters – galaxies: kinematics and dynamics – galaxies: structure – methods: numerical

1. Introduction

The Hubble (1926) sequence, the most widely used system of morphological classification of galaxies (Sandage & Tammann 1981), has served as a powerful framework for understanding galaxy evolution. Notwithstanding their myriad complexities, at the most fundamental level galaxies are principally distinguished by two dominant structural/dynamical components: a fast-rotating, flattened disk and a pressure-supported, spheroidal bulge. The relative light fraction of these two components, traditionally determined through photometric decomposition (e.g., Peng et al. 2002; Méndez-Abreu et al. 2008; Erwin 2015; Gao et al. 2019), establishes the galaxy type. Early-type galaxies are pure spheroids or bulge-dominated disk systems, whereas late-type galaxies are increasingly disk-dominated and even bulgeless. The processes that build up bulges and disks underlie the physical basis of the Hubble sequence.

Galaxies often comprise additional structures. For example, many nearby galaxies have a thick disk, which is both older and more metal-poor with respect to the thin disk (Dalcanton & Bernstein 2002; Yoachim & Dalcanton 2006; Comerón et al. 2011, 2014; Elmegreen et al. 2017). Meanwhile, the morphology of bulges comes in more than one flavor, ranging from highly spherically symmetric to flat (Andredakis & Sanders 1994; Andredakis et al. 1995; Courteau et al. 1996; Méndez-Abreu et al. 2010). Classical bulges are dynamically hot and largely featureless, likely the end-products of galaxy major mergers (Toomre 1977). The more flattened, rotationally supported pseudo bulges, an outgrowth of internal secular evolution, generally coexist with complex central structures (e.g., Erwin 2004; Kormendy & Kennicutt 2004). The Milky Way is a prototypical spiral that has several components, including a thin and a thick disk, a boxy/peanut-shaped bulge, a bar, a stellar halo, and a nuclear star cluster (see review by Bland-Hawthorn & Gerhard 2016). Boxy/peanut-shaped bulges are generally considered as the vertically thickened part of bars. Whether the Milky Way has a classical bulge is still uncertain, but it is unlikely to have a massive one (Shen et al. 2010; Debattista et al. 2017). The rich diversity of substructures observed among nearby galaxies imprints the formation and evolutionary history of galaxies. Accurate recognition and decomposition of these underlying substructures is essential.

The rapid development of integral-field spectroscopy has enabled galaxies to be classified by their internal kinematics (e.g., Emsellem et al. 2007, 2011; Cappellari et al. 2011a, 2011b). Early-type galaxies can be classified into slow and fast rotators (see the review of Cappellari 2016, and references therein), with fast-rotator early types forming a parallel sequence to spiral galaxies. Zhu et al. (2018b) made the first attempt to decompose observed galaxies based on their kinematics. Using the orbit-superposition Schwarzschild method (e.g., Schwarzschild 1979; Valluri et al. 2004; van den Bosch et al. 2008), they reconstructed stellar orbits for galaxies in the CALIFA survey (Sánchez et al. 2012), decomposing them into cold, warm, and hot components (Zhu et al. 2018a, 2018c). However, given the limited information that can be extracted from spectra, it is still very difficult to decompose observed galaxies in detail.

Numerical simulations are powerful tools for studying the formation and evolution of galaxy structures. In recent years, significant progress has been made in modeling star formation and stellar feedback, leading to increasingly realistic galaxies with reasonable bulge-to-disk ratios (Agertz et al. 2011; Guedes et al. 2011; Aumer et al. 2013; Stinson et al. 2013;

Marinacci et al. 2014; Roškar et al. 2014; Murante et al. 2015; Colín et al. 2016; Grand et al. 2017). The increase of simulation resolution has enabled us to generate galaxies with multiple structures that go much beyond the basic bulge+disk system, including vertical structures of disks (Brook et al. 2012; Ma et al. 2017; Navarro et al. 2018; Obreja et al. 2019), stellar halos (Cooper et al. 2010; Tissera et al. 2013; Pillepich et al. 2014; Elias et al. 2018; Monachesi et al. 2019), pseudo bulges (Guedes et al. 2013; Okamoto 2013), and bars (Algorry et al. 2017; Peschken & Łokas 2019).

Large-scale hydrodynamical cosmological simulations provide the opportunity to investigate the statistical properties of galaxies evolving in a fully cosmological context. Recent advances include Illustris (Genel et al. 2014; Vogelsberger et al. 2014a, 2014b), EAGLE (Crain et al. 2015; Schaye et al. 2015), and Horizon-AGN (Dubois et al. 2016). The IllustrisTNG simulations (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018, 2019; Pillepich et al. 2018a, 2019; Springel et al. 2018) can reproduce galaxies that successfully emulate plausible visual morphologies, thanks to an updated galaxy physics model (Weinberger et al. 2017; Pillepich et al. 2018b). The optical morphologies of galaxies in the TNG100 run (the highest-resolution version currently available at $z = 0$) are in good agreement with observations of nearby galaxies (Huertas-Company et al. 2019; Rodriguez-Gomez et al. 2019). The realism of the mock galaxies inspires confidence that the latest simulations can be used for detailed statistical study. With the aid of numerical simulations in which information is known in all six dimensions of phase space, we can investigate the intrinsic properties of galaxy structures, as well as track their formation physics and evolutionary history.

The structures of simulated galaxies can be identified through the kinematic properties of their constituent stars. Abadi et al. (2003) proposed a circularity parameter $\epsilon = J_z/J_c$, the ratio of the azimuthal angular momentum J_z and the maximum angular momentum J_c having the same binding energy E, that can separate effectively the spheroidal component from the disky component. In order to characterize different components in detail, Doménech-Moral et al. (2012) further introduced into consideration the binding energy E and the non-azimuthal angular momentum vector $\boldsymbol{J}_p = \boldsymbol{J} - \boldsymbol{J}_z$, where $\boldsymbol{J}$ is the total angular momentum vector of the stellar particle. These parameters identify the clustering of particles in kinematic phase space that corresponds to intrinsic structures of a galaxy. Obreja et al. (2016, 2018) replaced the k-means clustering algorithm used in Doménech-Moral et al. (2012) with an unsupervised machine-learning algorithm, the Gaussian mixture model (GMM). The use of a GMM reduces the errors caused by the mixtures of different structures via soft assignment of stars.

This study extends the application of GMM and develops a method to decompose simulated galaxies automatically. We test the method by applying it to five prototype galaxies from the TNG100 simulation (now also publicly available; see Nelson et al. 2019), and we discuss prospects for forthcoming applications using larger samples of simulated galaxies.

2. Dynamical Decomposition Method

Stars belonging to the same physical structure naturally cluster in their kinematic phase space. The method of Doménech-Moral et al. (2012) and Obreja et al. (2018) offers a promising framework to decompose the complex internal structures of galaxies. Their method uses three-dimensional (3D) Gaussian distributions to represent structures identified through their kinematics. However, a few limitations still affect its application:

1. The number of structures is determined artificially. This not only opens the possibility of human bias, but also renders impractical implementation to large samples of galaxies from cosmological simulations.
2. Real galaxy structures may not follow simple single-Gaussian distributions. The distribution function of disks, possibly all structures in galaxies, are not single Gaussians. A given structure may be composed of more complex distribution functions. Moreover, any realistic, dynamic, evolving system inevitably contains some degree of finer substructure.

An automated method, such as the unsupervised machine-learning nature of GMM, is needed to explore this problem. To mitigate human bias, the number of Gaussian components should be inferred directly from data, and all galaxies must be treated with the same standard. The fits should be sufficiently detailed to resolve significant structures in galaxies, allowing the same structure to host more than one Gaussian component if necessary. Our fully automated methodology, auto-GMM, complies with all these requirements and is able to identify multiple kinematic structures. We do not ascribe any physical significance to each individual component, postponing to a later stage the interpretation of the components/sub-components and their association with known, observed structures.

2.1. 3D Kinematic Phase Space

For TNG100, we load the positions and velocities of all particles (including dark matter, gas, and star) within a selected subhalo. Then, the code of Obreja et al. (2018) is adopted to calculate the 3D kinematic phase space of j_z/j_c, j_p/j_c, and $e/|e|_{\max}$ of stars, which will be used as inputs to GMM. The quantities j_z, j_p, j_c, and e are the specific J_z, J_p, J_c, and E, respectively. The origin of the coordinate coincides with the galaxy center, which is defined as the minimum of the gravitational potential. The z-axis of the galaxy is oriented perpendicularly to the outer disk. The average angular momentum vector is calculated by stars whose radii are between 2.1 kpc (three times the softening radius of stars) and 0.1 times the virial radius. Then, the azimuthal term of the angular momentum j_z can be easily decomposed from j_p. In order to estimate j_c and e, the code recalculates the gravitational potential of the halo, under the assumption that the halo is isolated. This assumption is generally well satisfied, unless the galaxy is undergoing significant accretion, which is fairly rare at low redshifts. All of the dark matter, stellar, and gaseous masses are included to recalculate the gravitational potential. The quantity $|e|_{\max}$ is the absolute value of the energy of the most bound stellar particle in the halo. Thus, $e/|e|_{\max}$ describes how tightly bound or centrally concentrated a particle is. It is a dimensionless parameter that gives a typical value across all galaxy masses. Only particles with $j_z/j_c \in [-1.5, 1.5]$, $j_p/j_c \in [0, 1.5]$, and $e \in [-1, 0]$ are considered in dynamical decomposition, consistent with the criteria used in Obreja et al. (2018). These criteria reject interlopers that are particles with clearly different kinematics from the galaxy.

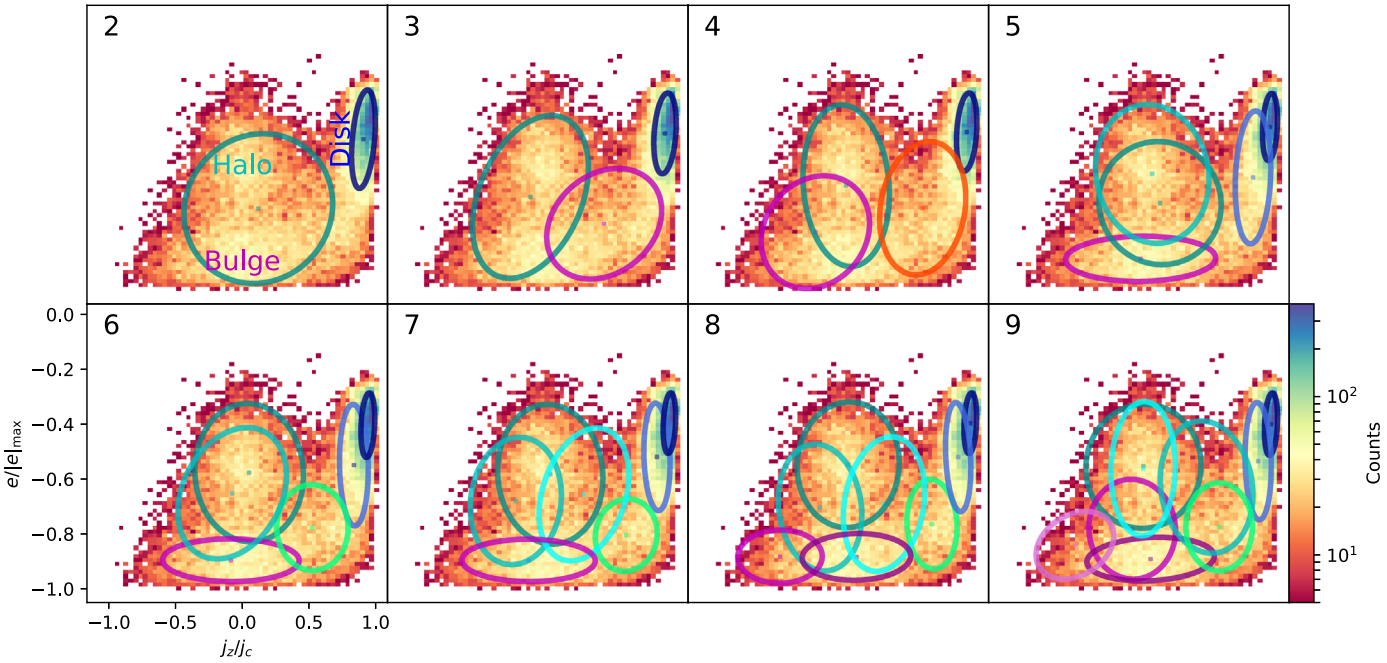

Figure 1. Distribution of j_z/j_c vs. $e/|e|_{\max}$ of a typical galaxy. Three components, likely corresponding to a disk, a bulge, and a halo, are visible. The color bar indicates the number of stellar particles in each bin. We fit the distribution of particles with increasing numbers of Gaussian components, from 2 to 9, as labeled in the upper-left corner of each panel. The overlaid ellipses represent 63% confidence regions of the Gaussian components found by the GMM fits. The fit improves by adding more components. Both the bulge (magenta ellipses) and disky components ($j_z/j_c > 0.7$; blue ellipses) are well fitted using $n_c = 5$–8. With the increase of n_c, the halo breaks up into multiple substructures (cyan ellipses).

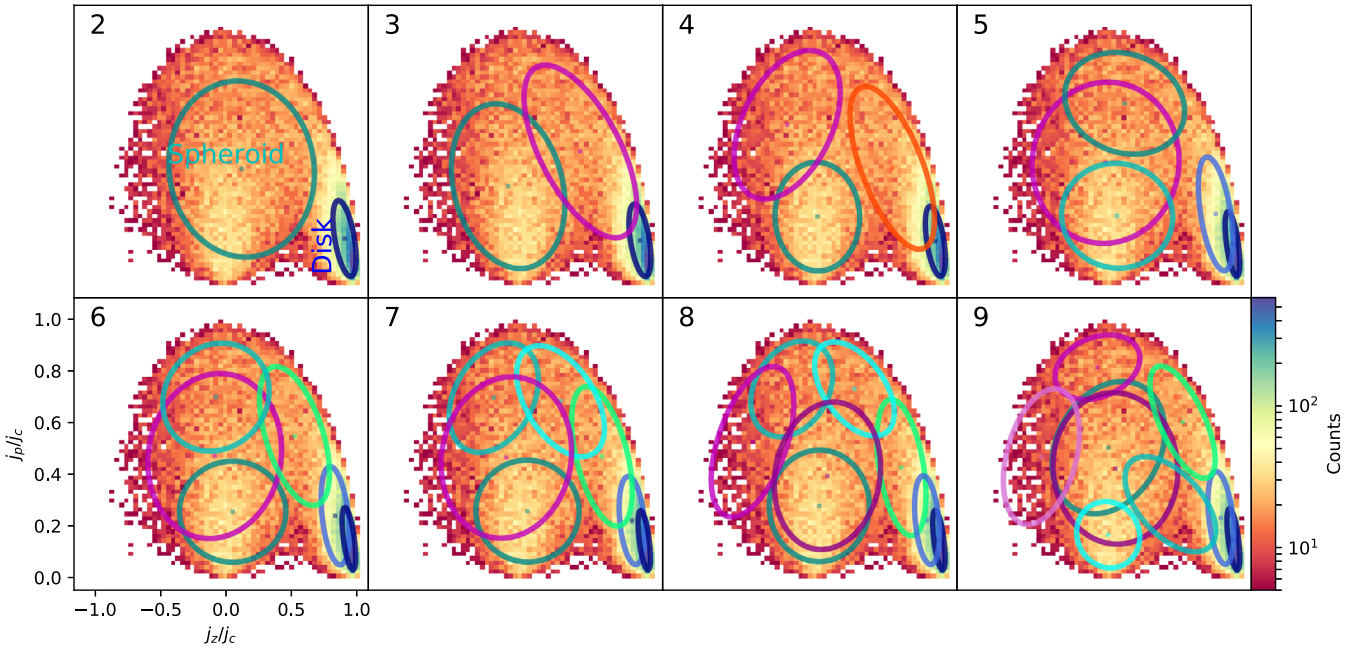

Figure 2. Distribution of j_z/j_c vs. j_p/j_c of the same galaxy shown in Figure 1. A spheroidal and a disky component are clearly visible. The color bar indicates the number of stellar particles in each bin. We fit the distribution of particles with increasing numbers of Gaussian components, from 2 to 9, as labeled in the upper-left corner of each panel. The Gaussian components found by GMM are represented by ellipses, using the same color scheme as in Figure 1.

2.2. Gaussian Mixture Models

Unsupervised machine-learning algorithms can be used to cluster data points into different groups. The `PYTHON` language (Pedregosa et al. 2011) offers several clustering methods. As suggested by Obreja et al. (2018), GMM is suitable for finding structures in the kinematic phase space of j_z/j_c, j_p/j_c, and $e/|e|_{\max}$. In the updated `scikit-learn` package, the old GMM module is replaced by the `GaussianMixture` module. Each Gaussian component is a triaxial ellipsoid in kinematic phase space. In order to maximize the likelihood in the parameter space, an expectation-maximization algorithm iterates until the default criterion is satisfied, returning a matrix of probabilities. Each data point has a probability array of how likely it is that it belongs to a certain component.

As an example, Figures 1 and 2 illustrate the kinematic phase space of a galaxy with distinct spheroidal and disky structures.

Three components are clearly visible in the j_z/j_c versus $e/|e|_{\max}$ diagram (Figure 1), namely a compact and slow-rotating spheroid or bulge, a diffuse and slow-rotating spheroid or halo, and a fast-rotating disk (details about the identification of kinematic and morphological structures are provided in Section 3.2). By contrast, only two components are clearly seen in the j_z/j_c versus j_p/j_c diagram (Figure 2). This is because both spheroidal components, dominated by random motions, have a wide range of j_p/j_c.

We fit the kinematic phase space with GMM, varying the number of Gaussian components n_c from 2 to 9. In each case, the fit is performed 10 times with different initializations by setting the keyword `n_init = 10`. We emphasize that running enough initializations is very important to obtain a stable fit. All initial parameters are generated with the k-means algorithm.

Figures 1 and 2 use colored ellipses to represent the 63% confidence ellipse of each Gaussian distribution obtained by the GMM fit. A disky and a spheroidal component can be roughly represented by setting $n_c = 2$ (top-left panel), but the kinematics of the galaxy are apparently more complex than such a simple, conventional bulge+disk decomposition. As expected, the fit improves by adding more components, but the data clearly become overfit when $n_c \geqslant 9$. The kinematics of this galaxy are well reproduced with $n_c = 5$–8. Both the bulge (magenta ellipses) and disky components ($j_z/j_c > 0.7$; blue ellipses) are well fitted, while the halo breaks up into multiple substructures (cyan ellipses) with increasing n_c. The halo and bulge components show no distinct separation in the j_z/j_c versus j_p/j_c diagram. Thus, components identified purely in the j_z/j_c versus j_p/j_c plane are not as robust as those decomposed in the j_z/j_c versus $e/|e|_{\max}$ plane.

2.3. Bayesian Information Criterion

Instead of artificially choosing n_c, we derive a modified version of the Bayesian information criterion (BIC) for selecting n_c in GMM. The BIC, developed by Schwarz (1978) and widely used in analysis of clustering data, allows the user to infer an approximate posterior distribution over the parameters of a Gaussian mixture distribution. Its formal definition is

$$\mathrm{BIC} = -2n \cdot \ln(\widehat{L}) + k \cdot \ln(n), \tag{1}$$

where $\ln(\widehat{L})$ is the average log-likelihood of a given data set, n is the number of data points, and k is the number of free parameters to be estimated. Because the geometry of each Gaussian distribution is fully relaxed by allowing a free 3D covariance matrix, GMM adds 10 extra free parameters (1 weight, 3 means, and 6 covariances) for each additional Gaussian component (i.e., $k = 10n_c$). The BIC is a decreasing function of $\ln(\widehat{L})$ and an increasing function of k. Hence, the second term on the right-hand side of Equation (1) is a penalty for the number of parameters introduced in the fit and serves to limit overfitting. A model having a smaller BIC is preferred, which implies either fewer free parameters or a better fit.

The mean BIC of each data point is

$$\widehat{\mathrm{BIC}}(n, n_c) = \frac{\mathrm{BIC}}{n} = -2\ln(\widehat{L}(n_c)) + \frac{10n_c \ln(n)}{n}. \tag{2}$$

This form is more meaningful, as $\widehat{\mathrm{BIC}}$ quantifies how good a model is for each single stellar particle. However, it is not completely independent of n. We vary n_c from 2 to 15. Because n is generally rather large ($\gtrsim 10^5$), the penalty term is $\lesssim 0.01$, estimated from the case of $n = 10^5$ and $n_c = 10$. As a consequence, we cannot see a clear minimum $\widehat{\mathrm{BIC}}$; instead, $\widehat{\mathrm{BIC}}$ approaches an asymptotic value that changes little for $n_c > 10$. Additionally, as suggested in Section 2.2, using more than 10 Gaussian components in the fit is not well motivated physically. We define

$$\Delta\widehat{\mathrm{BIC}} = \widehat{\mathrm{BIC}} - \widehat{\mathrm{BIC}}_{\min}, \tag{3}$$

where $\widehat{\mathrm{BIC}}_{\min} = \sum_{n_c=11}^{15} \widehat{\mathrm{BIC}}(n_c)/5$ is the mean value of $\widehat{\mathrm{BIC}}(n_c > 10)$. $\Delta\widehat{\mathrm{BIC}}$ of every galaxy asymptotically reaches ~ 0. The number of components can be chosen as the minimum value that satisfies $\Delta\widehat{\mathrm{BIC}} < C_{\mathrm{BIC}}$, where C_{BIC} is our criterion for a reasonable GMM model. The choice of C_{BIC} will be discussed in the following section.

Our approach of combining GMM with BIC, which we call auto-GMM, takes advantage of the unsupervised nature of GMM and allows n_c to be inferred objectively and automatically from the data, with no additional assumptions imposed.

3. Application of Auto-GMM to Prototype Galaxies from `IllustrisTNG`

Auto-GMM allows us to decompose galaxies automatically and efficiently, making it a powerful tool for large data sets. In order to test the efficiency of this method, we apply it to prototype galaxies at redshift 0 from the TNG100.

3.1. C_{BIC} Inferred from the `IllustrisTNG` Galaxies

The criterion C_{BIC} is the only parameter that needs to be chosen artificially when auto-GMM is used. A proper C_{BIC} can be inferred statistically from the large sample of galaxies in `IllustrisTNG`. To ensure that the galaxies have meaningful, well-resolved structures, we only use galaxies with stellar masses that exceed $10^{10}\,M_\odot$, which corresponds to $>10^4$ stellar particles. For each star, we specify the parameter $\kappa_{\mathrm{rot}} = v_\phi^2/v^2$, which measures the relative importance of its kinetic energy in ordered rotation. Then, the average value of this quantity for each galaxy, which gives an indication of its morphology and kinematics, is $K_{\mathrm{rot}} = \sum_i m_i \kappa_{i,\mathrm{rot}}/M_\star$ (Sales et al. 2010), where m_i represents the mass of particle i and $M_\star$ is the total stellar mass of the system. More massive galaxies become increasingly dominated by random motions, such that $K_{\mathrm{rot}} \approx 0.3$ for $M_\star \gtrsim 10^{11}\,M_\odot$ (Figure 3; top panel). The mass ratio of spheroids f_{sph}, estimated by summing up stars with $\kappa_{\mathrm{rot}} < 0.5$, likely increases with increasing $M_\star$ (Figure 3; bottom panel). Both K_{rot} and f_{sph} are kinematic indicators of the morphology of the galaxies. All the parameters above are calculated using the stars of radius <30 kpc. In Figure 3, all galaxies of stellar mass $\geqslant 10^{10}\,M_\odot$ are included, but only unbarred galaxies satisfying $K_{\mathrm{rot}} \geqslant 0.5$ are selected for inferring the C_{BIC} of disk galaxies. We regard $K_{\mathrm{rot}} = 0.5$ as the criterion to separate elliptical and disk galaxies.

A massive, long bar complicates the kinematic decomposition (see discussion in Section 3.4). D. Zhao et al. (2019, in preparation) find that a significant fraction of local disk galaxies in the TNG100 have formed a bar. We only focus on unbarred galaxies here, in order to obtain a clean result. The sample of unbarred galaxies is selected using their maximum

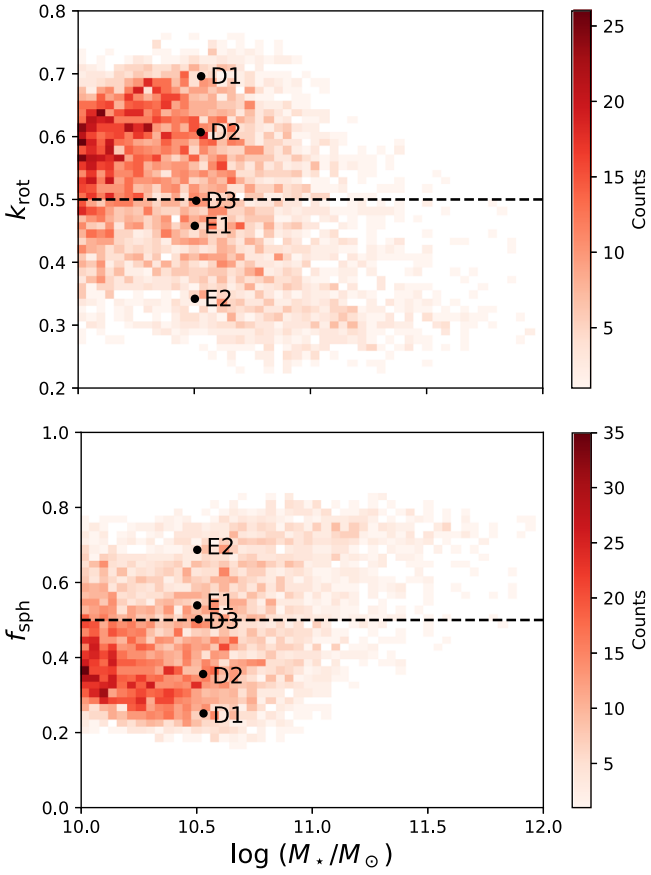

Figure 3. Fraction of the kinetic energy in ordered rotation, $K_{\rm rot}$, of the stellar particles (top) and the fraction of the stellar mass in the spheroidal component (bottom), $f_{\rm sph}$, as a function of stellar mass, $M_\star$, for the TNG100 galaxies. Here all 6503 galaxies, including barred galaxies, are shown. Five prototypes are marked with solid dots, classified into three disk galaxies (D1, D2, D3) and two ellipticals (E1, E2) with the criterion $K_{\rm rot} = 0.5$ (dashed lines). The color bar represents the number of galaxies in each bin. A total of 2994 unbarred disk ($K_{\rm rot} \geqslant 0.5$) galaxies are used for further analysis in Sections 3.1 and 3.2.

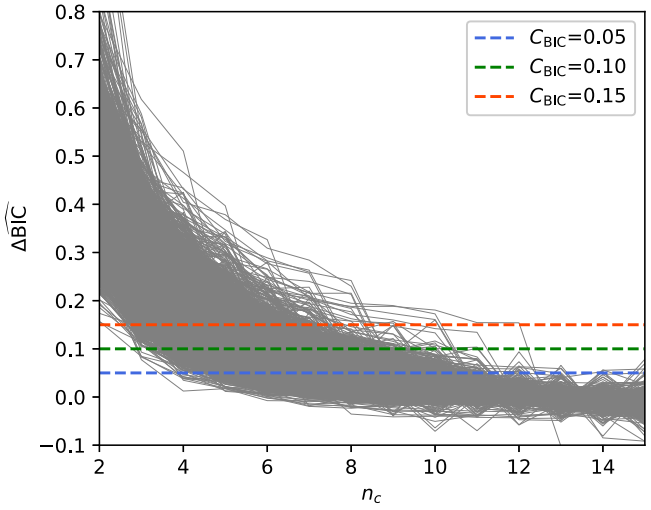

Figure 4. The $\widehat{\Delta{\rm BIC}}$ profiles as a function of n_c. All unbarred galaxies of stellar mass $>10^{10}\,M_\odot$ are included. We only exclude the elliptical galaxies, which are largely dominated by random motions ($j_z/j_c < 0.2$). The horizontal dashed lines mark different positions for the criterion $C_{\rm BIC}$ with values 0.05, 0.1, and 0.15.

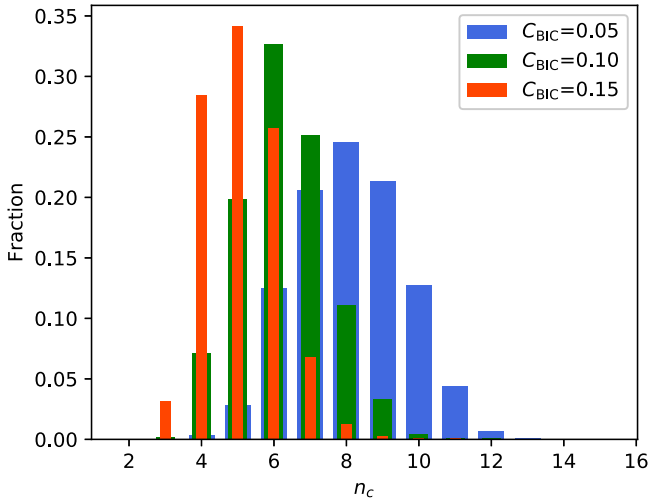

Figure 5. Distribution of the number of components n_c chosen by criterion $C_{\rm BIC} = 0.05$, 0.1, and 0.15 for the same galaxies shown in Figure 4.

ellipticity obtained from isophotal analysis of face-on images. Following standard convention (e.g., Marinova & Jogee 2007), a galaxy is considered unbarred if the maximum ellipticity is less than 0.25. We obtain a total of 2994 unbarred disk galaxies. This selected sample of unbarred disk galaxies is expected to have regular disky and spheroidal structures.

The $\widehat{\Delta{\rm BIC}}$ profiles of the selected galaxies are shown in Figure 4. We vary the criterion $C_{\rm BIC}$ from 0.05 (blue dashed line) to 0.15 (red dashed line); the corresponding distribution of n_c obtained with each $C_{\rm BIC}$ is shown in Figure 5. It is apparent that $C_{\rm BIC} = 0.05$ gives an unreasonable number of components ($n_c \approx 8$–11) for most galaxies. Both $C_{\rm BIC} = 0.1$ and 0.15 yield a reasonable number of components ($n_c \approx 4$–8). In general, $C_{\rm BIC} = 0.1$ results in one or two more components compared to $C_{\rm BIC} = 0.15$.

3.2. Intrinsic Structures Found in Unbarred Disk Galaxies from `IllustrisTNG`

A large library of GMM components in unbarred disk galaxies is built up by applying auto-GMM to TNG100. Then we need to associate these components to structures with which we are familiar from observations. Visual classification is not feasible for a large sample of galaxies. One reasonable way to classify GMM components automatically is by setting appropriate criteria on the mean values of j_z/j_c, j_p/j_c, and $e/|e|_{\rm max}$ of each component. The components belonging to the same structure should have similar properties, and hence should also cluster in kinematic phase space. Here we define the kinematic phase space of $\langle j_z/j_c\rangle$, $\langle j_p/j_c\rangle$, and $\langle e/|e|_{\rm max}\rangle$ as the mass-weighted mean values of j_z/j_c, j_p/j_c, and $e/|e|_{\rm max}$, respectively, of each Gaussian component.

The 2D histogram of mean circularity $\langle j_z/j_c\rangle$ versus mean rescaled energy $\langle e/|e|_{\rm max}\rangle$ of all components of the unbarred disk galaxies is shown in the left panel of Figure 6. There are four clear, distinguishable clusters that are likely to correspond to intrinsic structures. We can easily classify the components into spheroidal and disky structures by setting a threshold circularity criterion $\langle j_z/j_c\rangle = 0.5$ (thick dashed line). The spheroidal components can be classified further into bulges and halos by the criterion $\langle e/|e|_{\rm max}\rangle = -0.75$ (horizontal

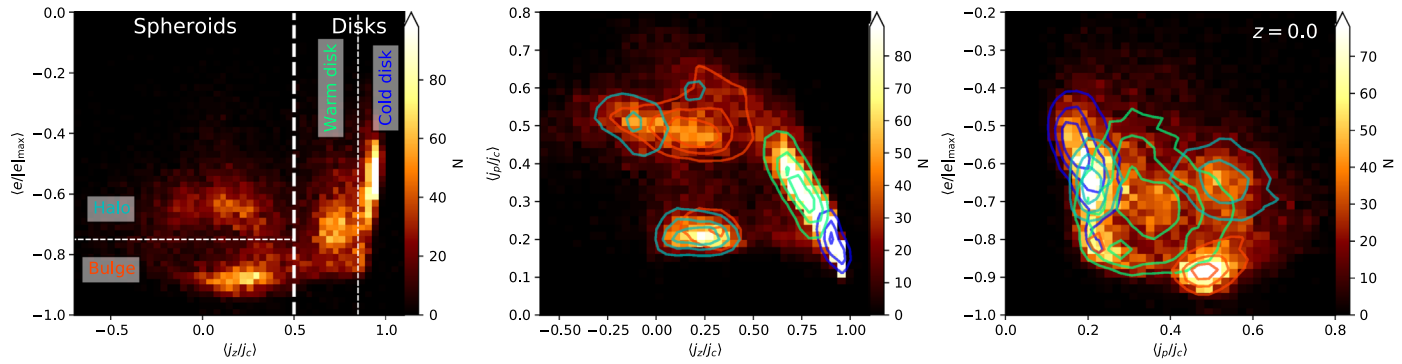

Figure 6. Kinematic phase space of all components of the unbarred disk galaxies from TNG100. The quantities $\langle j_z/j_c\rangle$, $\langle j_p/j_c\rangle$, and $\langle e/|e|_{\rm max}\rangle$ are the mean values of each Gaussian component found by auto-GMM with $C_{\rm BIC} = 0.1$. The color bar indicates the number of components in each bin. Four distinguishable clusters emerge in the diagram of $\langle j_z/j_c\rangle$ vs. $\langle e/|e|_{\rm max}\rangle$ (left panel): cold disk (blue), warm disk (green), halo (cyan), and bulge (red). The criteria adopted for this classification are marked with dashed lines. In the right two panels, we overlay the contours of these four kinds of structures on the map of number counts using the same color. The contours at levels 0.2, 0.4, and 0.6 are shown.

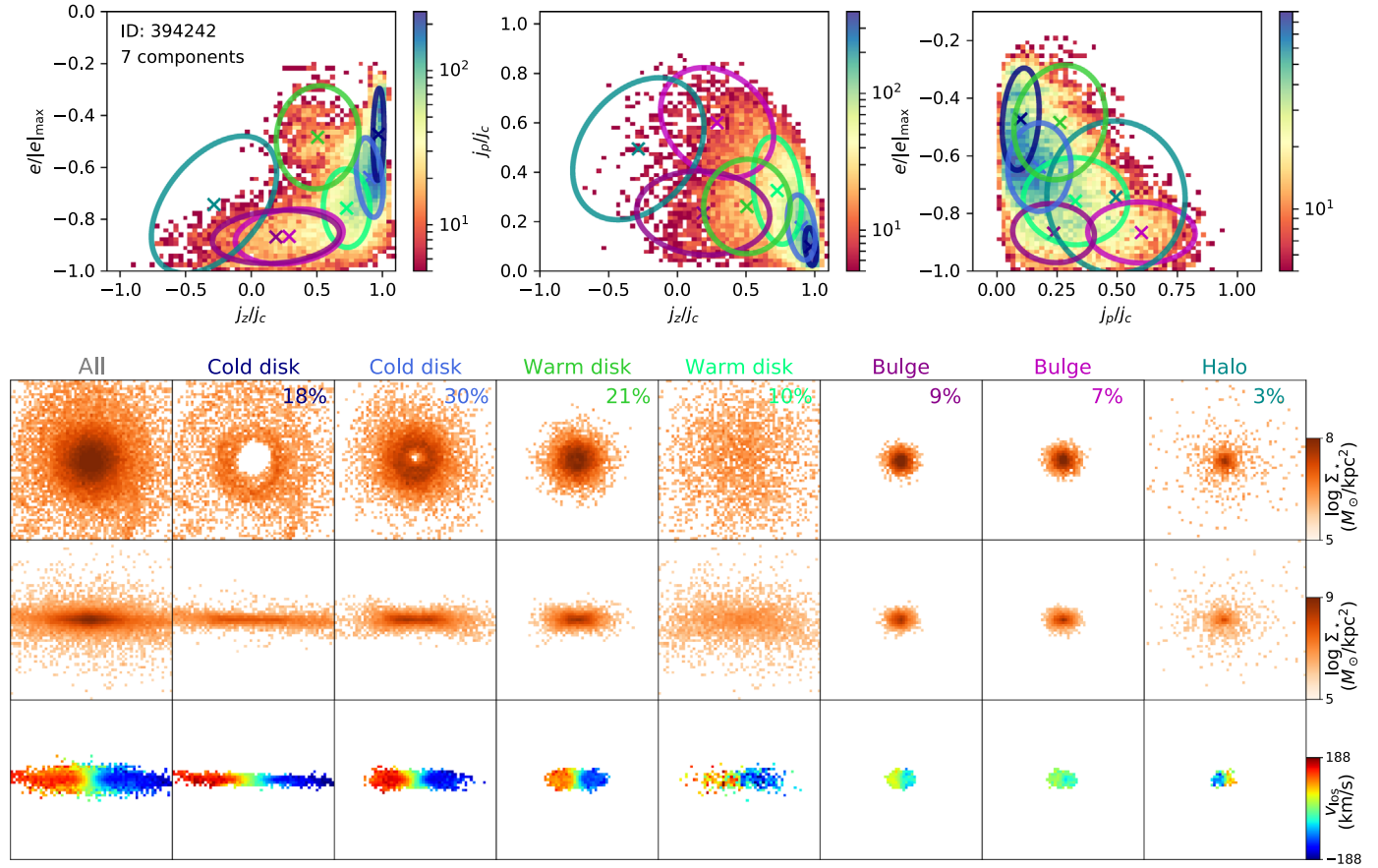

Figure 7. Model D1. The top row shows the kinematic phase space of j_z/j_c, j_p/j_c, and $e/|e|_{\rm max}$ and the Gaussian components found using auto-GMM. The TNG100 ID and the number of components are labeled in the first panel. The log-scale color bars of the top panels show the number of stars per bin in phase space. Seven components are found by auto-GMM using $C_{\rm BIC} = 0.1$. Their 63% confidence ellipses are overlaid, whose corresponding means are marked with crosses. The bottom three rows show the face-on and edge-on surface density distributions and the edge-on line-of-sight velocity distribution, respectively, of each component. These components are titled according to visual classification, and their corresponding mass fractions are labeled, using the same colors as those of the ellipses in the top row. For the line-of-sight velocity distribution, only bins having more than five particles are shown. The dimensions of the x and y axes are 60 kpc.

dashed line), while the disky components can be classified into cold disks and warm disks by $\langle j_z/j_c\rangle = 0.85$. Here j_p/j_c is not used in the classification, as it generally has quite a broad distribution for spheroids. To some extent, this is reasonable because spheroidal components may be composed of stars moving on highly radial orbits and in misaligned rotating orbits. The above strategy directly uses the data to statistically classify the components found by auto-GMM.

The middle and right panels of Figure 6 plot the contours of the number distribution of the components. The four series of contours correspond to the four kinds of structures. Both the cold and warm disks also cluster well in $\langle j_p/j_c\rangle$, while halos and

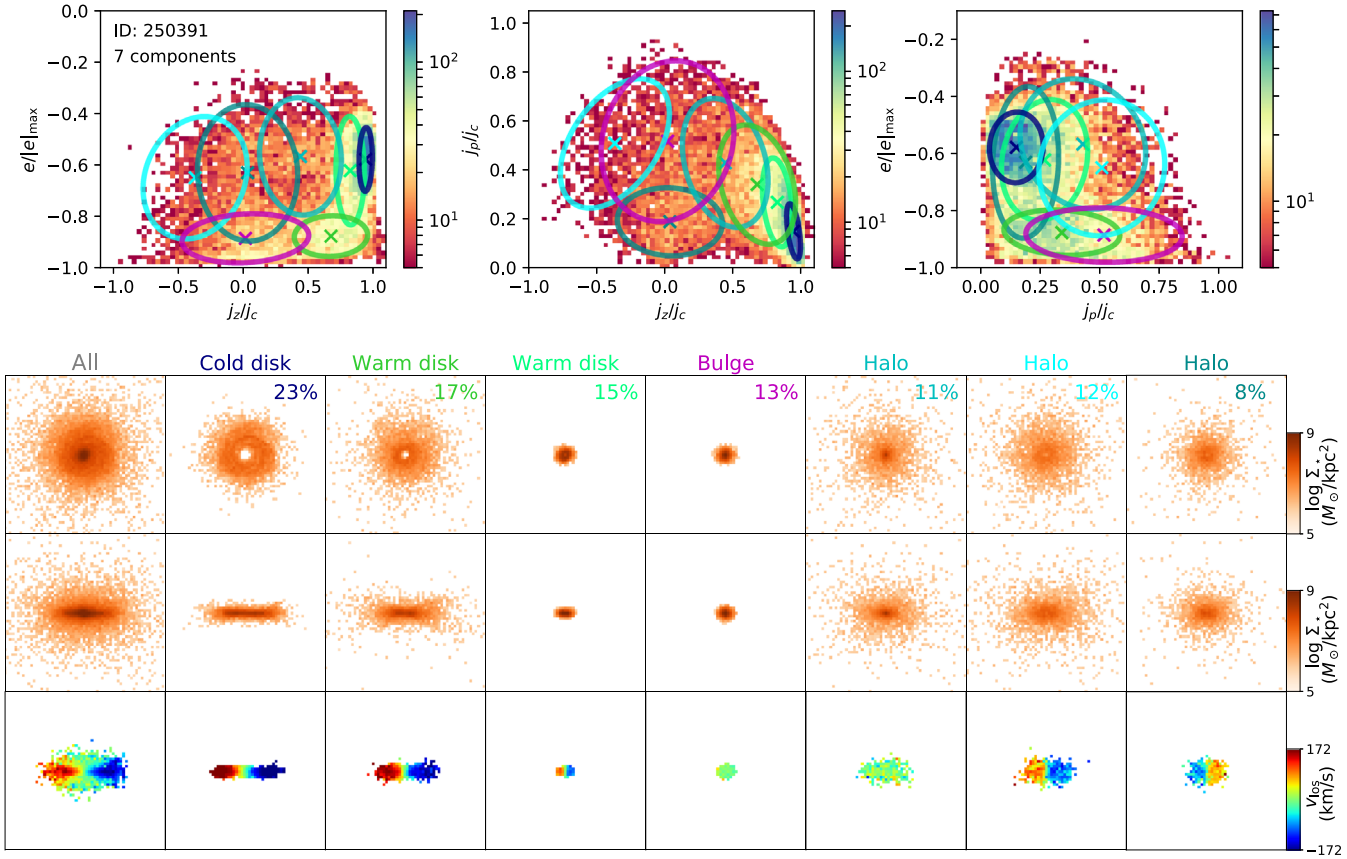

Figure 8. Model D2. Seven components are found by auto-GMM. The figure uses the same conventions as Figure 7.

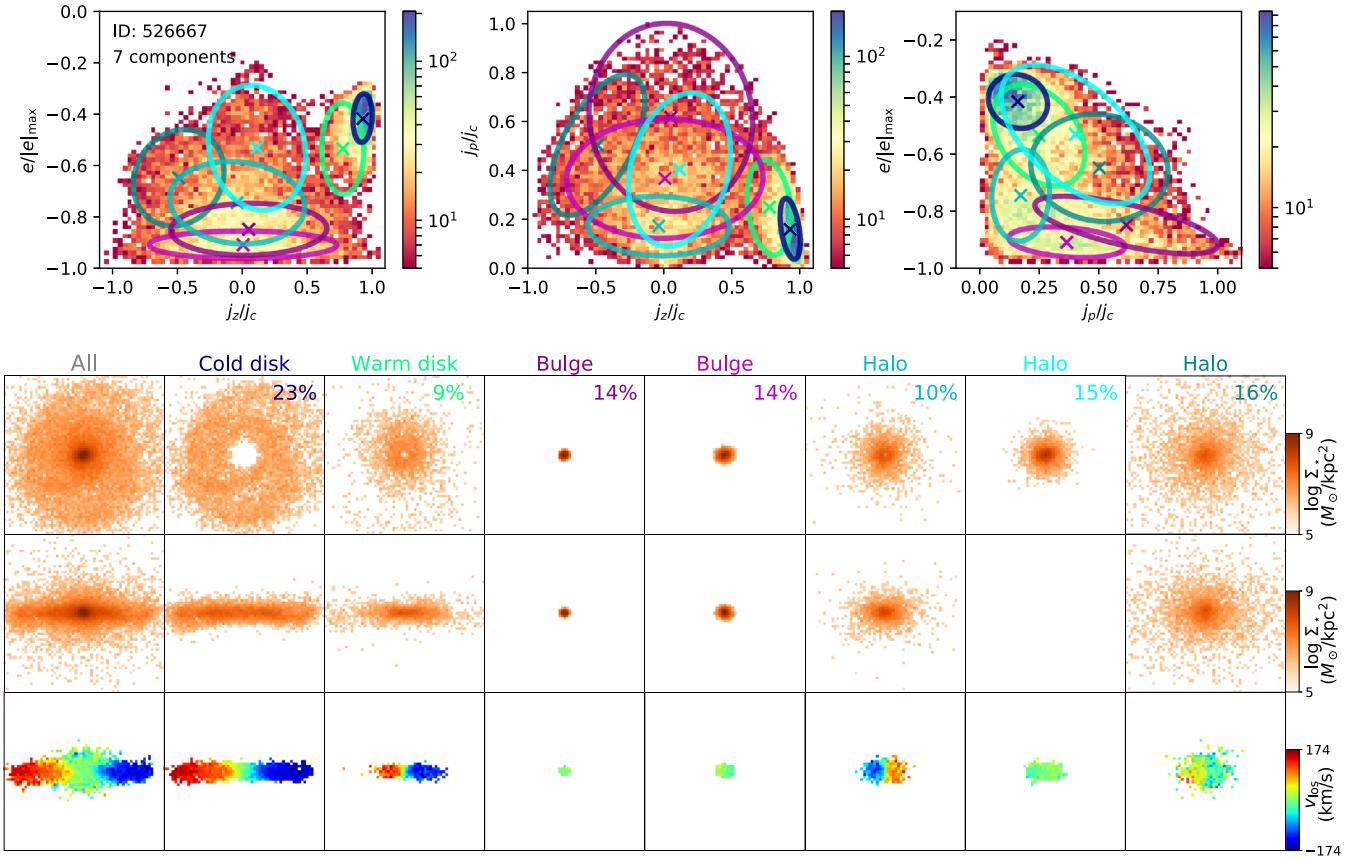

Figure 9. Model D3. Seven components are found by auto-GMM. The figure uses the same conventions as Figure 7.

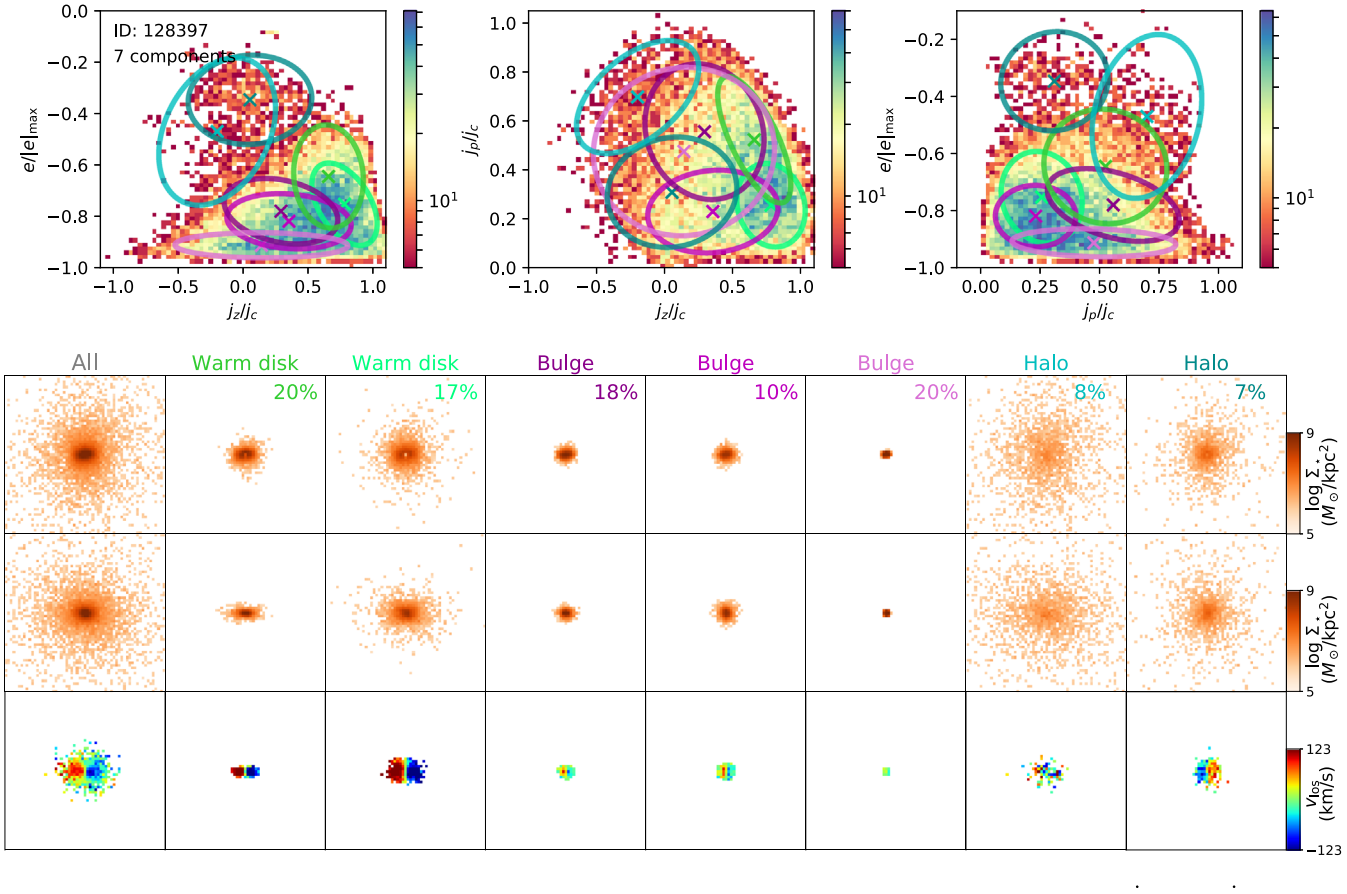

Figure 10. Model E1. Seven components are found by auto-GMM. The figure uses the same conventions as Figure 7.

bulges have two sub-groups of $\langle j_p/j_c \rangle \approx 0.5$ and $\sim$0.2, respectively. For spheroids, the components with $\langle j_p/j_c \rangle \approx 0.2$ are dominated by radial motions, while those with $\langle j_p/j_c \rangle \approx 0.5$ have significant, but misaligned, rotation.

We have demonstrated that the statistical results can be used to objectively infer the intrinsic structures of galaxies. The application of auto-GMM not only can decompose galaxies but also classify components in a completely automatic way. A few illustrative examples are shown in the next section.

3.3. Examples of Auto-GMM Fits

We choose a few prototype galaxies with diverse morphological and kinematic properties to test the performance of auto-GMM. In all cases, we adopt $C_{\rm BIC} = 0.1$. The five prototypes—three disk galaxies (D1, D2, D3) and two ellipticals (E1, E2)—all have the same stellar mass ($10^{10.5}\,M_\odot$) but cover a range of $K_{\rm rot}$ ($\sim$0.35–0.7) and $f_{\rm sph}$ ($\sim$0.25–0.75).

Figures 7–11 show the fits for models D1, D2, D3, E1, and E2, respectively. The first row shows diagnostic plots of j_z/j_c, j_p/j_c, and $e/|e|_{\rm max}$, with 63% confidence ellipses of all Gaussian components overlaid. The crosses mark their means. The 3D kinematic phase space of the five prototypes are well fit. From the second to the fourth row, we show, respectively, the face-on surface density, the edge-on surface density, and the line-of-sight velocity distribution for the edge-on view. Based on their properties, we classify the best-fit components into the structural families presented in Section 3.2: cold disk, warm disk, bulge, and halo. Their corresponding mass fractions are also given. Note that each identified structure can contain more than one component (e.g., two cold disk components in D1; two bulge components in D3), and we do not ascribe any particular interpretation to the physical nature of such substructures here.

Models D1 (Figure 7) and D2 (Figure 8) are largely dominated by disky structures. Components likely associated with cold disks and warm disks contribute 79% to the total stellar mass of D1 and 55% to the total stellar mass of D2. Only a small fraction of the mass in D1 arises from spheroidal components that we attribute to a bulge and halo. The two distinct "cold disk" components seen in the j_z/j_c versus $e/|e|_{\rm max}$ plot might share the same origin, with one portion being slightly dynamically hotter than the other, or they might originate from different gas accretion events. The two substructures of the bulge in D1 have similar compactness and weak rotation, differing principally only in their non-azimuthal angular momentum j_p. Such a difference is ignored in our classification, and they are considered as substructures of the same bulge.

Model D3 clearly has much more massive spheroidal components (Figure 9). Its disky components, including a cold (23%) and a warm (9%) disk, contribute only about one-third of the total stellar mass. The spheroidal components are impressively prominent. A clear pattern of rotation is still evident in the kinematic phase plot of D3. By contrast, the diagram of j_z/j_c versus $e/|e|_{\rm max}$ for model E1 is much more irregular and exhibits far fewer meaningful features (Figure 10). Violent mergers may have erased much of the substructure. Some mild rotation still exists, with $K_{\rm rot} \simeq 0.45$, arising mostly from an intermediate-scale disky structure.

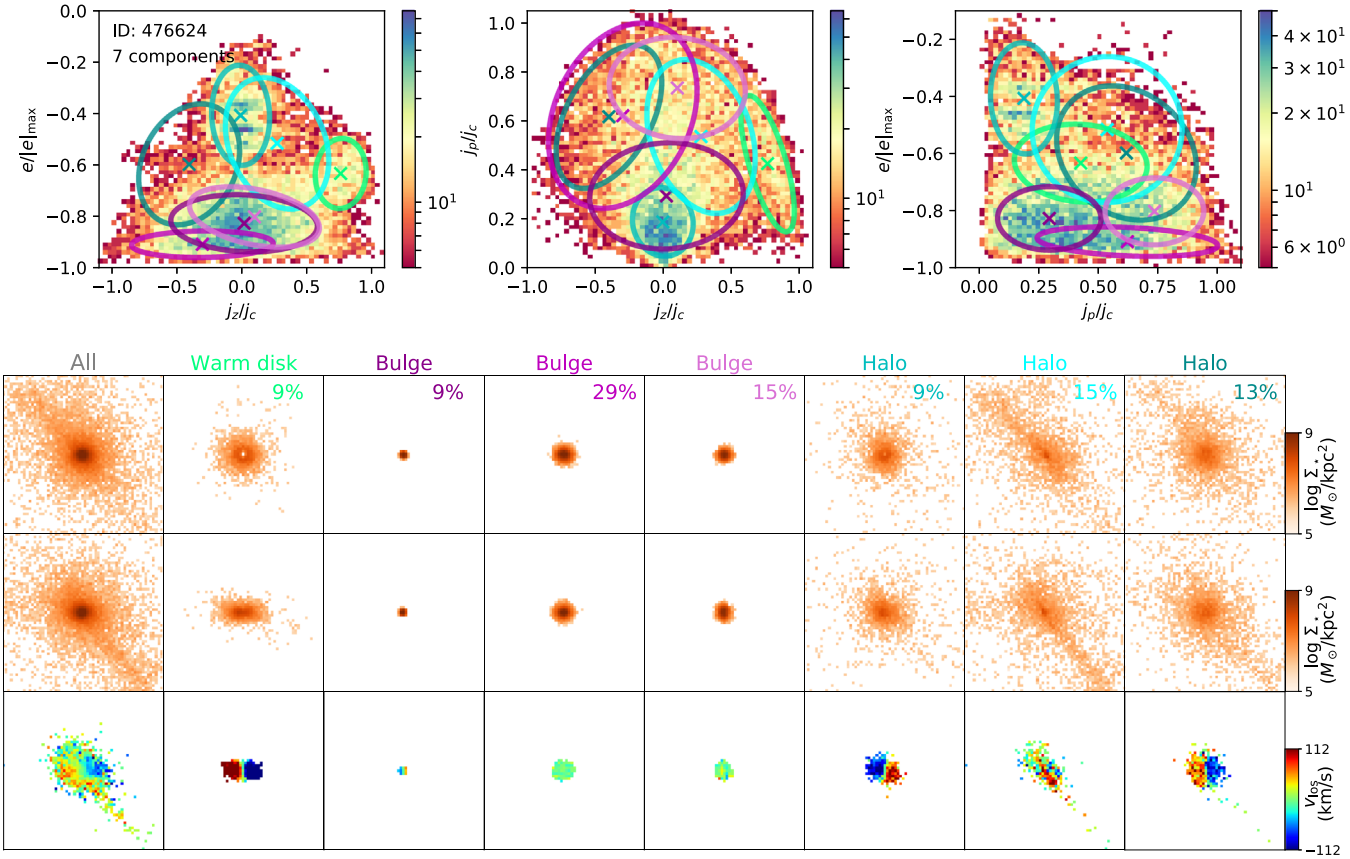

Figure 11. Model E2. Seven components are found by auto-GMM. The figure uses the same conventions as Figure 7.

Model E1 resembles moderate-mass ellipticals, which typically have disky isophotes and moderate rotation (e.g., Kormendy et al. 2009). Note that models E1 and D3 actually have similar rotation. However, D3 still maintains a clear disky morphology while E1 is quite spheroidal. E1 might be regarded as a lenticular galaxy given its mild rotation.

Cases with even lower values of $K_{\rm rot}$ are largely dominated by random motions. Model E2 is a typical elliptical galaxy with extremely weak rotation ($K_{\rm rot} \simeq 0.35$). The pattern of E2's kinematic phase space is more regular than that of E1. The bulges and halos classified by the criteria from the sample of unbarred disk galaxies correspond to a compact nuclear component and a diffuse envelope, respectively. A tidal tail due to a recent minor merger is still visible in the halo. The differences between E1 and E2 indicate that they may have experienced different assembly histories. Auto-GMM is also able to decompose typical elliptical galaxies, such as E2. However, many elliptical galaxies have a featureless kinematic phase space (e.g., E1). There is no proper way to model a featureless distribution even with multiple Gaussians. Thus, physically meaningless multiple Gaussians will be used to recover the data. Care is required in applying the auto-GMM method to elliptical galaxies.

It is worth emphasizing that the relation between the structures decomposed by kinematics and those from morphological observations is still unclear. The morphologies of the structures found by auto-GMM here are roughly consistent with our expectations of thin disks, thick disks, bulges, and halos. However, there are some essential differences. On the one hand, the bulge defined in kinematics is the tightly bound/compact part of spheroids, while the halo is the diffuse part. Halos do contribute to the central density, which is indistinguishable in observations of most of external galaxies. Thus, the inner part of kinematic halos will be considered as part of (classical) bulges in observations. Whether bulges and halos are formed in the same way, namely through mergers, is beyond the scope of this paper. On the other hand, warm disks may be related to thick disks and pseudo bulges in observations, and may have formed via very diverse pathways. Forthcoming papers will statistically investigate the properties and evolution of the structures identified here.

3.4. The Failure of Auto-GMM Fits in Barred Galaxies

Particles moving on bar orbits have complex kinematics that are unlikely to be well described by the phase space of j_z/j_c, j_p/j_c, and $e/|e|_{\rm max}$. Figure 12 shows an example of an auto-GMM fit of a typical barred galaxy from `IllustrisTNG`. At a given radius, particles moving on bar orbits rotate more slowly compared with those on circular orbits, and j_z/j_c decreases gradually with decreasing $e/|e|_{\rm max}$. As a consequence, bar particles significantly pollute the components having moderate rotation, such as warm disks. At the same time, bar particles with $j_z/j_c < 0.3$ may also influence significantly the kinematic decomposition of slowly rotating components, probably even bulges. Under this circumstance, the mass of the bulge is clearly overestimated. Bar particles do not cluster well in this kinematic phase space, as shown in the first row of Figure 12. They instead drive significant mixture in the kinematic phase space between disks and spheroids. Therefore, the auto-GMM method fails to reliably decompose barred galaxies.

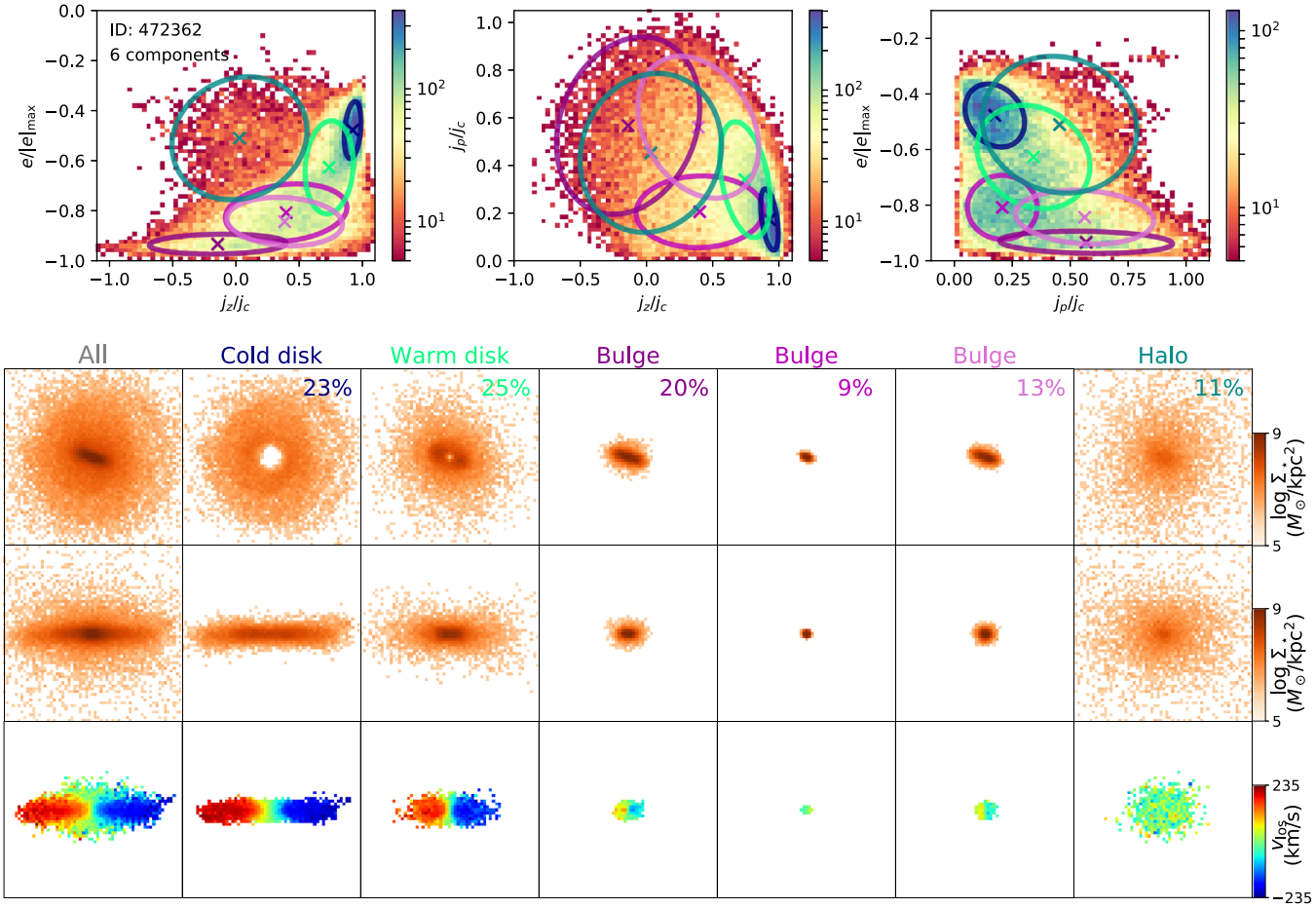

Figure 12. Galaxy hosting a strong bar. Six components are found by auto-GMM with $C_{\rm BIC} = 0.1$. The figure uses the same conventions as Figure 7.

4. Summary

We have described an automated method, auto-GMM, that generalizes the GMMs to decompose the stellar kinematics of simulated unbarred galaxies. A modified version of the BIC is used to infer the optimal number of statistically significant Gaussian components to fit the data.

We demonstrate that the simulated galaxies display rich substructures that can be identified and decomposed effectively by auto-GMM in the kinematic phase space of the stellar particles. Each substructure is a 3D Gaussian component. The substructures belonging to the same structure also cluster in the diagram of the mean circularity versus the compactness (rescaled energy) of the Gaussian components. Taking advantage of a large sample of galaxies in the cosmological simulation `IllustrisTNG`, four kinds of intrinsic structures are identified: cold disks, warm disks, bulges, and halos. While the present study does not ascribe any rigorous physical interpretation to the decomposed individual components, we illustrate the power of the auto-GMM method to isolate features that can be plausibly associated with morphological components (cold disk, warm disk, bulge, halo) traditionally associated with structures in the Hubble sequence of galaxies.

Our proposed method is automated, fast, and effective. It is a powerful tool to analyze a large data set of galaxies from cosmological simulations to gain insights into the origin and nature of galaxy structure. In forthcoming work, we will statistically investigate the properties of structures in thousands of galaxies from IllutrisTNG. We hope that the results can help interpret observations and provide more insight into the formation and evolution of real galaxies.

This work was supported by the National Science Foundation of China (11721303) and the National Key R&D Program of China (2016YFA0400702). M.D. is also supported by the grants "National Postdoctoral Program for Innovative Talents" (#8201400810) and "Postdoctoral Science Foundation of China" (#8201400927) from the China Postdoctoral Science Foundation. D.Y.Z. and J.S. acknowledge the support by the Peking University Boya Fellowship. V.P.D. was supported by STFC Consolidated grant ST/R000786/1. The TNG100 simulation used in this work, one of the flagship runs of the IllustrisTNG project, has been run on the HazelHen Cray XC40-system at the High Performance Computing Center Stuttgart as part of project GCS-ILLU of the Gauss centers for Supercomputing (GCS). The authors thanks all the IllustrisTNG team for making the IllustrisTNG data available to us prior to the public release. We thank the anonymous referee for valuable comments. We also thank Dandan Xu and Aura Obreja for constructive discussions. This work is highly supported by the High-performance Computing Platform of Peking University, China. The analysis was performed using `Pynbody` (Pontzen et al. 2013).

ORCID iDs

Min Du https://orcid.org/0000-0001-9953-0359
Luis C. Ho https://orcid.org/0000-0001-6947-5846

Dongyao Zhao https://orcid.org/0000-0001-8592-7910
Victor P. Debattista https://orcid.org/0000-0001-7902-0116
Lars Hernquist https://orcid.org/0000-0001-6950-1629

References

Abadi, M. G., Navarro, J. F., Steinmetz, M., & Eke, V. R. 2003, ApJ, 597, 21
Agertz, O., Teyssier, R., & Moore, B. 2011, MNRAS, 410, 1391
Algorry, D. G., Navarro, J. F., Abadi, M. G., et al. 2017, MNRAS, 469, 1054
Andredakis, Y. C., Peletier, R. F., & Balcells, M. 1995, MNRAS, 275, 874
Andredakis, Y. C., & Sanders, R. H. 1994, MNRAS, 267, 283
Aumer, M., White, S. D. M., Naab, T., & Scannapieco, C. 2013, MNRAS, 434, 3142
Bland-Hawthorn, J., & Gerhard, O. 2016, ARA&A, 54, 529
Brook, C. B., Stinson, G. S., Gibson, B. K., et al. 2012, MNRAS, 426, 690
Cappellari, M. 2016, ARA&A, 54, 597
Cappellari, M., Emsellem, E., Krajnović, D., et al. 2011a, MNRAS, 413, 813
Cappellari, M., Emsellem, E., Krajnović, D., et al. 2011b, MNRAS, 416, 1680
Colín, P., Avila-Reese, V., Roca-Fàbrega, S., & Valenzuela, O. 2016, ApJ, 829, 98
Comerón, S., Elmegreen, B. G., Salo, H., et al. 2014, A&A, 571, A58
Comerón, S., Knapen, J. H., Sheth, K., et al. 2011, ApJ, 729, 18
Cooper, A. P., Cole, S., Frenk, C. S., et al. 2010, MNRAS, 406, 744
Courteau, S., de Jong, R. S., & Broeils, A. H. 1996, ApJL, 457, L73
Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937
Dalcanton, J. J., & Bernstein, R. A. 2002, AJ, 124, 1328
Debattista, V. P., Ness, M., Gonzalez, O. A., et al. 2017, MNRAS, 469, 1587
Doménech-Moral, M., Martínez-Serrano, F. J., Domínguez-Tenreiro, R., & Serna, A. 2012, MNRAS, 421, 2510
Dubois, Y., Peirani, S., Pichon, C., et al. 2016, MNRAS, 463, 3948
Elias, L. M., Sales, L. V., Creasey, P., et al. 2018, MNRAS, 479, 4004
Elmegreen, B. G., Elmegreen, D. M., Tompkins, B., & Jenks, L. G. 2017, ApJ, 847, 14
Emsellem, E., Cappellari, M., Krajnović, D., et al. 2007, MNRAS, 379, 401
Emsellem, E., Cappellari, M., Krajnović, D., et al. 2011, MNRAS, 414, 888
Erwin, P. 2004, A&A, 415, 941
Erwin, P. 2015, ApJ, 799, 226
Gao, H., Ho, L. C., Barth, A. J., & Li, Z.-Y. 2019, ApJS, 244, 34
Genel, S., Vogelsberger, M., Springel, V., et al. 2014, MNRAS, 445, 175
Grand, R. J. J., Gómez, F. A., Marinacci, F., et al. 2017, MNRAS, 467, 179
Guedes, J., Callegari, S., Madau, P., & Mayer, L. 2011, ApJ, 742, 76
Guedes, J., Mayer, L., Carollo, M., & Madau, P. 2013, ApJ, 772, 36
Huertas-Company, M., Rodriguez-Gomez, V., Nelson, D., et al. 2019, MNRAS, 489, 1859
Kormendy, J., Fisher, D. B., Cornell, M. E., & Bender, R. 2009, ApJS, 182, 216
Kormendy, J., & Kennicutt, R. C., Jr. 2004, ARA&A, 42, 603
Ma, X., Hopkins, P. F., Wetzel, A. R., et al. 2017, MNRAS, 467, 2430
Marinacci, F., Pakmor, R., & Springel, V. 2014, MNRAS, 437, 1750
Marinacci, F., Vogelsberger, M., Pakmor, R., et al. 2018, MNRAS, 480, 5113
Marinova, I., & Jogee, S. 2007, ApJ, 659, 1176
Méndez-Abreu, J., Aguerri, J. A. L., Corsini, E. M., & Simonneau, E. 2008, A&A, 478, 353
Méndez-Abreu, J., Simonneau, E., Aguerri, J. A. L., & Corsini, E. M. 2010, A&A, 521, A71
Monachesi, A., Gómez, F. A., Grand, R. J. J., et al. 2019, MNRAS, 485, 2589
Murante, G., Monaco, P., Borgani, S., et al. 2015, MNRAS, 447, 178
Naiman, J. P., Pillepich, A., Springel, V., et al. 2018, MNRAS, 477, 1206
Navarro, J. F., Yozin, C., Loewen, N., et al. 2018, MNRAS, 476, 3648
Nelson, D., Kauffmann, G., Pillepich, A., et al. 2018, MNRAS, 477, 450
Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 2
Obreja, A., Dutton, A. A., Macciò, A. V., et al. 2019, MNRAS, 487, 4424
Obreja, A., Macciò, A. V., Moster, B., et al. 2018, MNRAS, 477, 4915
Obreja, A., Stinson, G. S., Dutton, A. A., et al. 2016, MNRAS, 459, 467
Okamoto, T. 2013, MNRAS, 428, 718
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266
Peschken, N., & Łokas, E. L. 2019, MNRAS, 483, 2721
Pillepich, A., Nelson, D., Hernquist, L., et al. 2018a, MNRAS, 475, 648
Pillepich, A., Nelson, D., Springel, V., et al. 2019, MNRAS, in press
Pillepich, A., Springel, V., Nelson, D., et al. 2018b, MNRAS, 473, 4077
Pillepich, A., Vogelsberger, M., Deason, A., et al. 2014, MNRAS, 444, 237
Pontzen, A., Roškar, R., Stinson, G. S., et al. 2013, pynbody: Astrophysics Simulation Analysis for Python, Astrophysics Source Code Library, ascl:1305.002
Rodriguez-Gomez, V., Snyder, G. F., Lotz, J. M., et al. 2019, MNRAS, 483, 4140
Roškar, R., Teyssier, R., Agertz, O., Wetzstein, M., & Moore, B. 2014, MNRAS, 444, 2837
Sales, L. V., Navarro, J. F., Schaye, J., et al. 2010, MNRAS, 409, 1541
Sánchez, S. F., Kennicutt, R. C., Gil de Paz, A., et al. 2012, A&A, 538, A8
Sandage, A., & Tammann, G. A. 1981, A Revised Shapley-Ames Catalog of Bright Galaxies (Washington, DC: Carnegie Institution)
Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, MNRAS, 446, 521
Schwarz, G. 1978, The Annals of Statistics, 6, 461
Schwarzschild, M. 1979, ApJ, 232, 236
Shen, J., Rich, R. M., Kormendy, J., et al. 2010, ApJL, 720, L72
Springel, V., Pakmor, R., Pillepich, A., et al. 2018, MNRAS, 475, 676
Stinson, G. S., Bovy, J., Rix, H.-W., et al. 2013, MNRAS, 436, 625
Tissera, P. B., Scannapieco, C., Beers, T. C., & Carollo, D. 2013, MNRAS, 432, 3391
Toomre, A. 1977, in Evolution of Galaxies and Stellar Populations, ed. B. M. Tinsley, R. B. G. Larson, & D. Campbell (New Haven, CT: Yale Univ. Observatory), 401
Valluri, M., Merritt, D., & Emsellem, E. 2004, ApJ, 602, 66
van den Bosch, R. C. E., van de Ven, G., Verolme, E. K., Cappellari, M., & de Zeeuw, P. T. 2008, MNRAS, 385, 647
Vogelsberger, M., Genel, S., Springel, V., et al. 2014a, MNRAS, 444, 1518
Vogelsberger, M., Genel, S., Springel, V., et al. 2014b, Natur, 509, 177
Weinberger, R., Springel, V., Hernquist, L., et al. 2017, MNRAS, 465, 3291
Yoachim, P., & Dalcanton, J. J. 2006, AJ, 131, 226
Zhu, L., van de Ven, G., Méndez-Abreu, J., & Obreja, A. 2018a, MNRAS, 479, 945
Zhu, L., van den Bosch, R., van de Ven, G., et al. 2018b, MNRAS, 473, 3000
Zhu, L., Ven, G. v. d., Bosch, R. v. d., et al. 2018c, NatAs, 2, 233